John von Neumann Institute for Computing

The NIC Complex Systems Research Group

Peter Grassberger

published in

NIC Symposium 2006 ,
G. Münster, D. Wolf, M. Kremer (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **32**, ISBN 3-00-017351-X, pp. 3-11, 2006.

http://www.fz-juelich.de/nic-series/volume32

The NIC Complex Systems Research Group

Peter Grassberger

John von Neumann Institute for Computing
Research Centre Jülich, 52425 Jülich, Germany
E-mail: p.grassberger@fz-juelich.de

We give an overview over the main research activities in our group during the last two years. It is mostly concentrated on two subjects: Efficient sequential Monte Carlo methods and pattern analysis. As regards sequential Monte Carlo algorithms, we mainly developed a fast algorithm for the simulation of lattice animals and lattice trees. This follows in spirit our PERM ("pruned-enriched Rosenbluth method") algorithm which was highly successful for polymers and other problems, but it involves a number of non-trivial departures from plain PERM. In the following, we will discuss results on various aspects of the animal problem obtained with this algorithm. As regards pattern analysis, our main long-term commitment is to EEG analysis for understanding epilepsy. One very important technique in this is independent component analysis (ICA), which allows quite generally to decompose a signal into weakly interdependent sources. One essential ingredient in ICA is an algorithm to estimate this interdependency with least bias. In our group we first developed a new estimator of mutual information (the main information-based measure of interdependency), and we then used this in two new classes of ICA algorithms. The first is a conventional algorithm called MILCA ("mutual information based least-dependent component analysis"), the other is a simulated annealing method using Markov chain Monte Carlo and is today the most efficient algorithm for problems with non-negative sources arising in spectral analysis.

1 Introduction

The John von Neumann Institute has two research groups. One is for high energy physics, the other for various problems of statistical physics and complex systems at the interface between physics and biology. In the present paper I want to give a brief review of the activities during the last two years when I was head of the group for complex systems.

As also during the previous years, our group was active in two main fields: Efficient Monte Carlo sampling by means of recursive sequential algorithms, and signal analysis. Although members of the group have also worked on other important problems, I will in the following concentrate on these two, and will present one achievement in each.

2 PERM Simulations for Lattice Animals

As regards Monte Carlo methods, one distinguishes Markov chain (Metropolis-Hastings) methods and sequential methods. It is fair to say that simulations in statistical physics use overwhelmingly the Metropolis strategy, and sequential algorithms have traditionally occupied niches like neutron transport theory[1], diffusion type quantum Monte Carlo methods[2], and some algorithms (called "static" in this field) for sampling polymer configurations[3–6]. But the last years have seen a resurgence of sophisticated sequential methods, mostly among statisticians[8], while physicists have still been more busy with more sophisticated algorithms of the Metropolis type.

Our group has made a major contribution with the PERM algorithm (for a review, see[7]), which is a recursively implemented ("depth first") version of sequential sampling with importance sampling and re-sampling[8]. Although this is a general purpose strategy and was applied to as different problems as reaction-diffusion systems[9], percolation[7], and sequence analysis[10], its main application was to polymer statistics. It allowed e.g. the first detailed test of the logarithmic corrections predicted for so-called Θ-polymers by the renormalization group[11], and provides today the most efficient algorithm for determining low energy configurations of lattice heteropolymers which provide toy models for protein folding[12].

During the last two years, we have applied the basic strategy of PERM to the simulation of lattice animals and lattice trees. Lattice animals (or "polyominoes", as they are also called by mathematicians[13]) are just connected clusters of sites on some regular lattice, similar to percolation clusters. But while the latter come with non-trivial statistical weights which are due to the fact that percolation clusters are created by randomly removing bonds or sites from a fully occupied lattice, the animal ensemble is defined so that all clusters with the same number of sites have the same weight. Thus understanding the animal problem essentially corresponds to counting the numbers of different cluster shapes. The interest for statistical physics derives on the one hand from the fact that lattice animals form the simplest model in the universality class of randomly branched polymers[14]. On the other hand, they are closely related to a number of other problems. A famous theorem, conjectured by Parisi and Sourlas[15] and proven much later[16], connects lattice animals in d dimensions to the Lee-Yang problem in $d-2$ dimensions. As a consequence, some critical exponents are known exactly in 3 and 4 dimensions, while other exponents are not known exactly even in 2 dimensions, because animals are not conformally invariant[17].

Efficient simulations of lattice animals have always posed a problem. For lattice trees (which are just animals with tree topology) a version of the pivot algorithm[18] is fairly efficient, but even that is much more cumbersome and slow than algorithms for percolation clusters or for unbranched polymers. For percolation clusters, in particular, one has very simple growth algorithms first given by Leath[19], which can be implemented either as breadth or as depth first[20].

Our PERM algorithm starts by growing slightly subcritical percolation clusters and re-weighs them as appropriate for the animal ensemble. It is crucial that this can be done while the cluster is still growing. So one can check whether the weight is just right, too small, or too large (as measured against the average over previous clusters; at the start, when there are not yet any previous clusters, every cluster is 'just right'). If it is too large, the cluster is "cloned", each clone receives half of the weight, and continues to grow independently of the other. If the weight is too small, the cluster is killed with probability 1/2 and the weight of the survivors is doubled. All this could be done with an explicit population of clusters as in an evolutionary (genetic) algorithm, but we found it more convenient to use a depth-first (stack oriented, recursive) implementation. Notice that here "depth-first" refers not to the way how an individual cluster grows, but to the way how the state space tree of different growth paths is sampled. Indeed, we found that for an efficient algorithm it is crucial to use a breadth-first growth algorithm.

The algorithm involves a number of other parameters and choices, all of which are crucial for efficiency, although very precise choices of the parameters are often not needed. For instance, we have to choose the control parameter of the percolation process with which

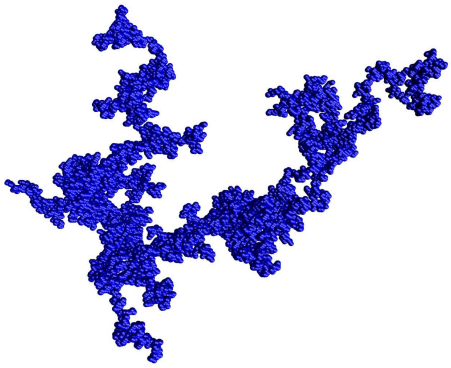

Figure 1. A typical 3-d lattice animal with 16000 sites on the bcc lattice. According to our simulations, there are about $10^{16776.0\pm0.4}$ different cluster shapes of this size.

we start, and we have to choose a "fitness function" which tells us when to clone and kill. All these details are described in[21]. A large typical animal (16,000 sites) is shown in Fig.1, animals of somewhat smaller sizes were simulated with high statistics in all dimensions between 2 and 9.

Among the results obtained with this algorithm is, first of all, the verification that the critical exponents are in agreement with the Parisi-Sourlas conjecture, with exponents as obtained by field theoretic methods for the Lee-Yang problem. For the exponent ν in 2 dimensions ($\langle R^2\rangle \sim N^{2\nu}$ where R is the radius of gyration and N is the number of sites) we obtain excellent agreement with exact enumeration results.

For clusters grafted to an attractive surface we find the expected phase transition between desorbed and adsorbed phases, and we verify partially a striking prediction by Janssen and Lyssy[22]: The cross-over exponent is exactly 1/2 in 3 and 4 dimensions. In $d = 2$, Janssen and Lyssy also conjectured a value 1/2, but with less theoretical justification. Indeed we find small deviations for $d = 2$ which seem to be significant.

In order to understand whether conformal invariance is replaced for lattice animals by some other simple property, we simulated clusters grafted to the apex of $2 - d$ cones and wedges (a cone is a wedge whose two sides are glued together)[23]. While the exponent ν is independent of the opening angle α of the wedge/cone, the entropic exponent θ does depend on α. For conformally invariant theories, this dependence is $\theta \propto 1/\alpha$. While we found this to be true in the limit $\alpha \to 0$, the dependence for large angles is linear, $\theta \propto \alpha$. In spite of the suggestive simplicity of this result, we have not found an easy explanation.

Finally, we have simulated collapse transitions due to attractive forces between the 'monomers' making up to clusters. There are two natural parameters for this attraction[24]. Correspondingly, one obtains a 1-dimensional collapse line in a 2-dimensional phase plot. Critical (bond) percolation is obtained with special choices of these two parameters and sits exactly on the phase transition line. Away from it, the exact location of the transition

line is not known, as are also the critical exponents associated with it. Even worse, also the topology of the phase diagram is still debated[24, 25]. Our simulations[26] gave very precise results in part of the phase diagram, but not in all. In particular, we still cannot clearly resolve the dispute about its topology.

3 Independent Component Analysis and Other Applications of Mutual Information

As said above, we have a long-standing commitment to understand the EEG of epilepsy patients. The main goals there are to predict seizures in advance (typically a few minutes), and to locate the epileptic focus precisely for later resection. The latter should be done preferentially only from data obtained during from seizure-free intervals. A recent review of the successes and difficulties encountered by our collaboration on this project is in[27].

One of the main features of epilepsy is the very strong EEC signal during seizures, which can only be generated by strong synchronization effects. One expects interdependencies between the EEC signals obtained from different regions in the brain to show some traces of this also when there is no seizure. Thus we started already rather early to study interdependence measures[28]. We did not use at that time mutual information (MI), theoretically the ideal measure because of its strong information theoretic background[29], because of the well-known problems to obtain unbiased MI estimates (for a recent discussion, see e.g. Ref. 30).

This changed when we found that a class of estimators, based on the k-th nearest neighbour method of Kozachenko and Leonenko[31] for estimating differential Shannon entropies for real-valued variables, has surprisingly good properties[32]. It has small statistical errors, is reasonably fast when implemented carefully, and seems to have zero bias when applied to random variables which actually are independent. The latter purely numerical observation (we have no proof for it) is very astonishing, and is particular useful for applications to independent component analysis (ICA)[33].

It was therefore mainly to the latter that we applied it up to now. ICA is based on the assumption that a set of simultaneously measured signals $x_i(t_k)$ are actually composed of more or less independent sources. In the simplest case, there is no additional noise (or, otherwise said, some of the sources are just noise), these sources are strictly independent, and the signals are obtained from them by linear superposition with a constant (i.e. time independent) and instantaneous *mixing matrix* $\mathbf{A}$. In "blind source separation", both $\mathbf{A}$ and the sources are assumed to be unknown, and the goal is to recover them from the statistics of the signals alone. Obviously, when the number of possible sources is not larger than the number of measured components, this is done by applying an instantaneous time-constant "demixing" matrix $\mathbf{W} = \mathbf{A}^{-1}$ such that the reconstructed sources $s_i(t) = (\mathbf{W}\mathbf{x}(t))_i$ are as independent as possible.

A crucial ingredient for any version of ICA is obviously a good measure of statistical dependencies. If one takes the simple linear correlation coefficients, then the problem can be solved by linear analysis, the result being just a principle component decomposition. Usually this is used as a first step in the analysis ("pre-whitening"), and ICA proper is obtained by using a more sophisticated (in general non-linear) measure of dependence. Very popular are various approximations to MI or higher order cumulants like skewness

Figure 2. Estimated independent components of the heart beat of a pregnant woman, using the public domain JADE algorithm. Traces nr. 5 and 6 are dominated by the heart beat of the fetus, but the signals are still quite noisy.

or kurtosis (e.g. in the JADE[34] and FastICA[33] algorithms), but also time-delayed linear correlations are used, as e.g. in the TDSEP algorithm[35].

In real world applications, the mixing might be not strictly linear or might involve some time delays, or the true sources might be not strictly independent. Thus we have to expect the reconstructed sources not to be strictly independent either. Our strategy is then to lump together sources whose MIs are above a given threshold into one multi-dimensional source. This is done by using a hierarchical clustering algorithm where MI is used as proximity measure[36]. Possible time delays can be incorporated easily by working with delay embeddings[37] familiar from nonlinear analysis of univariate signals. In order to see the contribution of any particular reconstructed (multi-dimensional) source to the measured signals, one turns off all other sources and applies the inverse of the reconstructed demixing matrix.

When implemented in a rather standard way, i.e. with pre-whitening and with an iterated deterministic gradient descent method for the minimization of MI, we called our algorithm MILCA ("MI based least dependent component analysis"). By applying it to various test cases, we invariably found it to be better than all competitors, with the exception of a recent method also based on k-th nearest neighbour entropy estimators[38]. The latter uses a rather time consuming trick which the authors called 'data augmentation'. When we included augmentation in our bag of tricks, our algorithm did even beat that method.

In Figs. 2 to 4 we show the more realistic case of the heart beat (ECG) of a pregnant woman. The original data[39] have 8 channels recorded for 5 seconds. In Fig. 2 we show the results obtained with a standard public domain algorithm, JADE[34]. We see that the heartbeat of the fetus is reasonably well separated from that of the mother and from sources which mainly consist of noise. When applying MILCA, we compare two versions: one without delay embedding, and one with three delay times, which blows the number of

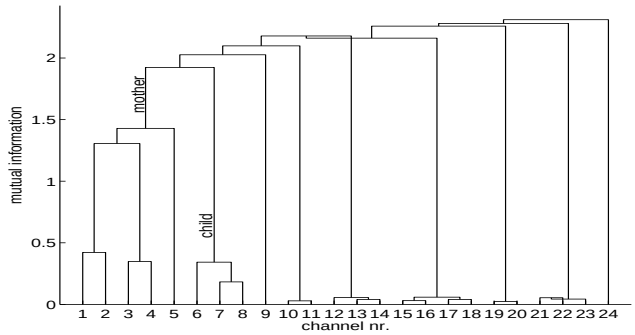

Figure 3. Dendrogram obtained from the dependencies between the 24 sources reconstructed with MILCA. Heights of each cluster correspond to the MI between its elements.

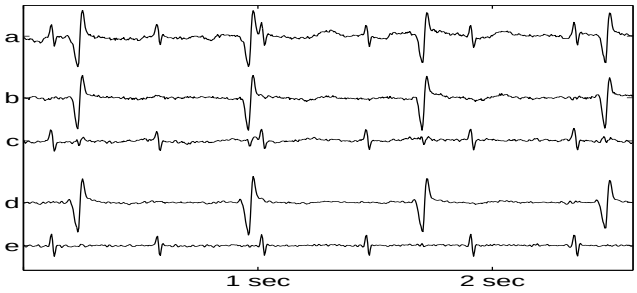

Figure 4. Short segment (a) of the original EC; (b) of the mother and (c) of the fetus contributions estimated with MILCA, but without delay embedding; and (d), (e) mother and fetus contributions obtained with delay embedding.

channels from 8 up to 24. In Fig. 3 we show the dendrogram obtained from the cluster analysis for the latter. We see two main clusters, obviously the heartbeat of the mother and the heartbeat of the fetus, while the other sources seem to be just noise. Suppressing everything except one of these two main clusters and applying the inverse demixing, we obtain the contributions of the two hearts to the full ECG. For better visibility we only show blow-ups of part of one single channel. This original recording of this channel is seen in Fig. 4a. In Figs. 4b and 4c we see the mother and fetus contributions, as estimates with MILCA without delay embedding. Finally, in Figs. 4d and 4e we see the results obtained with delay embedding. Even the heart beat of the fetus, which seems sometimes to be completely masked by the mother heart beat in the original data, is now recovered with extremely small noise.

Another large field of applications for blind source separation is spectroscopy. There one has typically a spatially (or temporally) inhomogeneous mixture of different substances, so that spectra obtained from different locations (at different times) correspond to different superpositions of the pure spectra. Here frequency plays the rôle that time had played before. Most ICA algorithms are based on the assumption that the signals are iid

(independent and identically distributed), i.e. that time correlations are absent. This is certainly a rather poor assumption for spectra. Also, spectra of different chemical substances often have large overlaps. Finally, many spectra not only are non-negative, they also have spectral regions where the intensity is close to zero. In such cases, looking just for decompositions with minimal MI might give rather poor results. In particular, this might lead to unphysical solutions with negative intensities. Indeed, negativities are in many cases introduced already in the prewhitening pre-processing step, and are then maintained during the main part of the ICA algorithm. There are methods to eliminate these negativities in a post-processing step, but then there is no control what happens to the independencies during this post-processing[40]. MILCA as described above performs thus for several test cases from the literature as good as typical state of the art codes, but not spectacularly good[40].

A more elegant way is to abandon prewhitening completely, and to perform an iterative demixing where non-negativity is preserved during each individual step. It turns out that a constraint greedy deterministic algorithm would then be very inefficient. Instead, we propose in[42] a Metropolis-type simulated annealing[41] algorithm with MI as cost function and with non-negativity as hard constraint. Results obtained with this SNICA ("stochastic non-negative independent component analysis") algorithm show that it gives better performance than any other algorithm proposed so far[42].

4 Discussion and Outlook

Obviously, there remains much to do. We started our work on ICA with the intention to apply it to the EEC, mainly of epilepsy patients. So far we have not done it, and since I now retired, I can only hope that others will continue with this work. There are of course many more applications of MI, in practically all fields of science. Whether our new MI estimators will become useful there, is a yet open question.

On the other side, there are also many more potential applications of sequential sampling algorithms. I mentioned already sequence analysis, where the generation of large samples of sequence pairs (or tuples, more generally) with prescribed statistics is needed for testing the significance of a found alignment. Another possible application is the generation of large random networks with given global properties (e.g. degree sequences)[43], which is also needed for null hypothesis testing.

Finally, there is a growing interest in the statistical community to apply Monte Carlo methods to Bayesian inference problems. This has obvious connections to nonlinear time series analysis, and would therefore tie together the two lines of research discussed in this review. Of course it was our long-time hope to bridge this gap sooner or later, but it seems now that this has to be done by other researchers.

Acknowledgments

It is a pleasure to thank my collaborators over all these years, in particular Walter Nadler, Hsiao-Ping Hsu, Ralph Andrzejak, Jochen Arnhold, Sergej Astakhov, Erwin Gerstner, Alexander Kraskov, Thomas Kreuz, Klaus Lehnertz, Vishal Mehra, Harald Stögbauer, and Lei Yang. I also want to thank the Forschungszentrum Jülich for the excellent support we had during this time.

References

1. H. Kahn, 'Use of Different Monte Carlo Sampling Techniques', in ed. H.A. Meyer, *Symposion on the Monte Carlo Method* (Wiley, New York 1956).
2. W. von der Linden, Phys. Rep. **220**, 53 (1992).
3. M.N. Rosenbluth *et al.*, J. Chem. Phys. **23**, 356 (1955).
4. F.T. Wall *et al.*, J. Chem. Phys. **30**, 634, 637 (1959).
5. S. Redner *et al.*, J. Phys. **A 14**, 2679 (1981).
6. T. Garel and H. Orland, J. Phys. **A 23**, L621 (1990).
7. P. Grassberger, Computer Physics Commun. **147**, 64 (2002).
8. J.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer Series in Statistics (Springer, New York 2001).
9. V. Mehra and P. Grassberger, Phys. Rev. E **65**, 050101 (2002); Physica D **168**, 244 (2002).
10. R. Bundschuh and P. Grassberger, work in progress.
11. P. Grassberger, Phys. Rev. E **56**, 3682 (1997).
12. H.-P. Hsu, V. Mehra, W. Nadler and P. Grassberger, Phys. Rev. E **68**, 021113 (2003).
13. S. Golomb, *Polyominoes: Puzzles, Patterns, Problems and Packings* (Princeton Univ. Press, Princeton, N.J. 1994).
14. T. C. Lubensky and J. Isaacson, Phys. Rev. A **20**, 2130 (1979).
15. G. Parisi and N. Sourlas, Phys. Rev. Lett. **46**, 871 (1981).
16. J.Z. Imbrie, J. Phys. A: Math. Gen. **37**, L137 (2004).
17. J.D. Miller and K. De'Bell, J. Physique I 3, 1717 (1993).
18. E J Janse van Rensburg and N Madras, J. Phys. A: Math. Gen. **25** 303 (1992).
19. P. Leath, Phys. Rev. B **14**, 5046 (1976).
20. R. Tarjan, SIAM J. Comput. **1**, 146 (1972).
21. H.-P. Hsu, W. Nadler, and P. Grassberger, J. Phys. A: Math. Gen. **38**, 775 (2005).
22. H. K. Jassen and A. Lyssy, J. Phys. A: Math. Gen. **25** L679 (1992); Europhys. Lett. **29**, 25 (1995).
23. H.-P. Hsu, W. Nadler, and P. Grassberger, Phys. Rev. E **71**, 065104 (2005).
24. S. Flesia, D.S. Gaunt, C.E. Soteros, and S.G. Whittington, J. Phys. A: Math. Gen. **27**, 5831 (1994).
25. M. Henkel and F. Seno, Phys. Rev. E **53**, 3662 (1996).
26. H.-P. Hsu and P. Grassberger, J. Stat. Mech. P06003 (2005).
27. F. Mormann *et al.*, Clinical Neurophys. **116**, 569 (2005).
28. J. Arnhold, P. Grassberger, K. Lehnertz, und C.E. Elger, Physica D 134, 419 (1999).
29. T.M. Cover and J.A. Thomas, *Elements of Information Theory* (Wiley, New York 1991).
30. C.J. Cellucci, A.M. Albano, and P.E. Rapp, Phys. Rev. E **71**, 066208 (2005).
31. L.F. Kozachenko and N.N. Leonenko, Probl. Inf. Transm. **23**, 95 (1987).
32. A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).
33. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (Wiley, New York 2001).
34. J.-F. Cardoso, Neural Computation **11**, 157 (1999).
35. F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller, IEEE Trans. Biomed. Eng. **49**, 1514 (2002).

36. A. Kraskov, H. Stögbauer, R.G. Andrzejak, and P. Grassberger, e-print q-bio.QM/0311039 (2003).
37. E. Ott, *Chaos in Dynamical Systems* (Cambridge Univ. Press, Cambrige 1993).
38. E.G. Learned-Miller and J.W. Fisher III, J. Machine Learning Res. **4**, 1271 (2003).
39. B.L.R. De Moor (ed), "Daisy: Database for the identification of systems", www.esat.kuleuven.ac.be/sista/daisy (1997).
40. S.A. Astakhov, H. Stögbauer, A. Kraskov, and P. Grassberger, e-print physics/0412029 (2004).
41. S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, Science **220**, 671 (1983).
42. S.A. Astakhov, H. Stögbauer, A. Kraskov, and P. Grassberger, preprint (2005), accepted for Angewandte Chemie.
43. Y. Chen, P. Diaconis, S.P. Holmes, and J.S. Liu, J. Amer. Statist. Assoc. **100**, 109 (2005).